Applying Markov Decision Process to Understand Driving Decisions Using Basic Safety Messages Data

Mohsen Kamrani[a,b,*], Aravinda Ramakrishnan Srinivasan[c,d], Subhadeep Chakraborty[c], Asad J. Khattak[a,*]

[a] *Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, TN 37996, United States*
[b] *Center for Urban Transportation Research, University of South Florida, Tampa FL 33620, United States (current affiliation)*
[c] *Department of Mechanical, Aerospace and Biomedical Engineering, University of Tennessee, Knoxville, TN 37996, United States*
[d] *Lincoln Center for Autonomous Systems, School of Computer Science, University of Lincoln, Lincolnshire, LN6 7TS, United Kingdom (current affiliation)*

[*] Corresponding author.
Email address: akhattak@utk.edu (Asad Khattak); mkamrani@vols.utk.edu (Mohsen Kamrani)

Abstract

While a number of studies have investigated driving behaviors, detailed microscopic driving data has only recently become available for analysis. Through Basic Safety Message (BSM) data from the Michigan Safety Pilot Program, this study applies a Markov Decision Process (MDP) framework to understand driving behavior in terms of acceleration, deceleration and maintaining speed decisions. Personally Revealed Choices (PRC) that maximize the expected sum of rewards for individual drivers are obtained by analyzing detailed data from 120 trips and the application of MDP. Specifically, this paper defines states based on the number of objects around the host vehicle and the distance to the front object. Given the states, individual drivers' reward functions are estimated using the multinomial logit model and used in the MDP framework. Optimal policies (i.e. PRC) are obtained through a value iteration algorithm. The results show that as the number of objects increases around a host vehicle, the driver prefer to accelerate in order to escape the crowdedness around them. In addition, when trips are segmented based on the level of crowdedness, increased levels of trip crowdedness results in a fewer number of drivers accelerating because the traffic conditions constrain them to maintaining constant speed or deceleration. One potential application of this study is to generate short-term predictive driver decision information through historical driving performance, which can be used to warn a host vehicle driver when the person substantially deviates from their own historical PRC. This information could also be disseminated to surrounding vehicles as well, enabling them to foresee the states and actions of other drivers and potentially avoid collisions.

Keywords: Driving Behavior, Markov Decision Processes, Basic Safety Messages, Multinomial Logit Model, Instrumented Vehicle Data, Automation

1. Introduction

The availability of detailed driving performance data provides new opportunities to investigate different aspects of driving behavior. Some of these aspects are exhibited through vehicle motion (e.g., speed, acceleration). In this paper, driving behavior is defined as instantaneous driving decisions in terms of acceleration, deceleration and maintaining constant speed, which vary based on contexts. Detailed models are available for the dynamics of vehicle components (Guzzella and Sciarretta, 2007, Kiencke and Nielsen, 2005). In some models, the driver is a feedback controller that seeks to achieve a particular control goal, such as tracking a reference (Burnham et al., 1974, Prokop, 2001). In other cases, the driver is represented by an autonomous system, often driven by a random process. For instance, (Macadam, 2003) proposes a model with linear and non-linear elements that include actuator saturation, slew-rate, and time delay, (Liu and Pentland, 1997) suggests a hidden Markov model, and (Cooper, 1991) proposes non-linear Autoregressive–moving-average (ARMAX) models. (Kiencke et al., 1999) introduces a hybrid driver model which consists of discrete modes and continuous control functions.

Driving behavior models (Toledo et al., 2007) generally describe vehicle movements in different traffic conditions. These prediction methods include speed, acceleration and lane changing models and are critical in microscopic traffic simulators. Other application areas where aggregate traffic flow characteristics are extracted from individual driving behavior, such as safety and capacity analysis, could also benefit from such models.

Early driving behavior models focused on car-following theory. These models explain the behavior of a following vehicle assuming it reacts to the lead vehicle's actions (Brackstone and McDonald, 1999, Rothery, 1992). Recently, the increased use of microscopic traffic simulation models has stimulated the development of general acceleration models and lane changing behavior studies. General acceleration models (Gipps, 1981, Yang and Koutsopoulos, 1996) define multiple driving regimes (e.g., free-flow, emergency) while considering different behaviors in each regime at various car-following types (e.g., reactive and non-reactive). For instance, drivers in the free-flow acceleration regime may focus on attaining their desired speed. In lane changing models, (Gipps, 1986, Salvucci et al., 2001) there are typically two components: considering and executing the lane change maneuver (Karan and Chakraborty, 2016, Mohammadi et al., 2019). More recently, car following models driven by trajectory data that incorporate other contributing factors (e.g., distraction, reaction time) and address driver heterogeneity, have been developed (Li et al., 2016, Ossen and Hoogendoorn, 2005, Ossen and Hoogendoorn, 2011, Farah and Koutsopoulos, 2014, Hoogendoorn et al., 2013, Toledo et al., 2007, Koutsopoulos and Farah, 2012, Papathanasopoulou and Antoniou, 2015, Hoogendoorn et al., 2011, Ossen et al., 2006, Choudhury et al., 2009).

Finally, some studies have used Inverse Reinforcement Learning to recover driver reward functions and, consequently, driving styles. For instance, Shimosaka et al. (2015) tried to predict driving behavior by considering multiple reward functions and the maximum entropy method. Similarly, Kuderer et al. (2015), studied the possibility of learning a driver's style and navigation behavior through demonstration. Another study analyzes a driver's car following behavior using Continuous Inverse Optimal Control (Hayeri et al., 2016). More recently, researchers have used Deep Reinforcement Learning (Ye et al., 2019) and Mixed Observable Markov Decision Process

(MOMDP) (Sezer, 2018) to learn driving behavior and decisions from simulation data.

This study benefits from the availability of BSM data, which allows the authors to explore driving behaviors from a different perspective. The aim is to extract drivers' personally revealed choices (PRCs) from their acceleration/deceleration profiles. By treating instantaneous driving decisions as the realization of an optimal policy in an MDP framework, it is possible to define states over time in terms of objects surrounding a vehicle. The framework derives a driver's value for different actions they take (i.e. acceleration, braking or maintaining speed), which are quantified in terms of accumulated discounted rewards. When the expected sum of the driver's rewards is maximized, their personally revealed choices can be inferred, given different states. MDP is well suited for this study because of the Markov property of instantaneous driving decisions and the stochasticity of their outcomes. A stochastic process has the Markov property if the probability of future states depends only upon the present state, not on the sequence of events that led to it, i.e., past states. MDP is a reasonable approach for providing a framework for modeling decisions, which in this case are the decisions of a driver to accelerate, decelerate or maintain speed. MDP discretizes time in a way that is consistent with the near-instantaneous decisions that drivers make. MDP can also account for outcome randomness; it takes the complex driving environment problem and breaks it down to a state-action structure, making the problem tractable. The methods applied in this paper can potentially form a foundation for human driver personally revealed choice extraction using field-collected empirical data. In the near future, self-driving cars will claim a substantial share of the roadways. For a more collaborative driving experience, these cars will need to coordinate and anticipate the action space and propensities of human drivers in their vicinity. Currently, most of the focus and resources are pooled towards developing the sensor-interpretation-planning-execution loop in connected and automated vehicles (CAVs), but an efficient way of modeling human drivers' propensities in traffic will pave the way for a more 'human-like' driving style. Moreover, knowledge about individual driving behaviors can be used to generate alerts and warnings for the driver of a host vehicle and be passed on for the purpose of improving safety.

From a methodological standpoint, the paper contributes by using high volume and diverse driving data to learn driving decisions. Specifically, reward and states are defined theoretically and real-life driving decisions are analyzed to explore their correlations with contextual factors i.e., proximity to surrounding objects. The study reveals the personal preferences of drivers in each state. Since states are defined based on the number of objects around the host vehicle and the distance to the front object, the MDP reveals PRCs based on the level of the crowdedness around the host vehicle by accounting for the Markov property of drivers' decisions and the stochasticity of their outcomes.

2. Methodology

A dynamic traffic problem where state transitions can happen spontaneously due to the action of other vehicles and situations not under the control of the driver can potentially be at odds with the MDP assumption of stationary policy process. However, even with that amount of uncertainty, we as rational drivers still plan our trajectories and take actions according to our (often non-optimal)

plans/policies. We posit that human drivers base these decisions on a planning process that uses historically learnt "standard" state-transition probabilities and their relation to traffic density, etc. In fact, we go one step further and impose exceptions to these standard probabilities based on locally relevant information. In this paper, we capture how historically learnt information (through large volumes of aggregated data) can be used to come up with a plan/policy for multiple steps of driving. MDP is a suitable candidate for modeling such processes. In the application, the short-term departures from non-stationarity are neglected in exchange for long term converged probability values. This study aims to understand short-term behavior, driver decisions of acceleration, deceleration, and maintaining constant speed are conceptualized using the Markov Decision Process (MDP) (Bellman, 1954). A specific structure needs to be imposed on the MDP framework that models driver behaviors in terms of different maneuvers, which is explained below.

2.1. Markov Decision Processes

The Markov Decision Process, according to (Bellman, 1954) is defined by a set of states ($s \epsilon S$), a set of all possible actions (a ϵ A), a transition function ($T(s, a, s^{'})$), a reward function (R(s)), and a discount factor (γ). To make the model mathematically tractable, the discount factor is restricted to $0 < \gamma < 1$. In this study, the tuple $\{S, A, T, R, \gamma\}$ represents an MDP. The agent (e.g. decision maker, driver, etc.) takes an action based on the current state and an intuitively determined policy. The outcome of this action is stochastic and is parameterized as the transition probability function. Having transitioned to a new state, the decision maker accumulates some reward associated with that state. The process of decision-making continues infinitely. Generally, the term "Markov" implies that the future and the past states are independent given the present state. Specifically, for an MDP, this means that the future outcomes depend only on the current state and performed action. A policy $\pi(s)$ is a prescription for the action to be taken, given a current state. The total expected discounted reward for the agent following the policy is defined as (Sutton and Barto, 1998):

$$V^{\pi}(s) = R(s) + \gamma \sum_{s'} T(s, \pi(s), s')\, V^{\pi}(s') \tag{1}$$

This equation quantifies how valuable the state (s) is under the policy($\pi(s)$).

Likewise, the value or utility of a state-action pair $\left(Q^{\pi}(s, a)\right)$ is given by the following equation (Sutton and Barto, 1998):

$$Q^{\pi}(s, a) = R(s) + \gamma \sum_{s'} T(s, a, s')\, V^{\pi}(s') \tag{2}$$

Policy $\pi(s)$ is referred to as the optimal policy $\pi^{*}(s)$ *(*hereafter PRC*)*, if it satisfies the Bellman's optimality equation (Bellman, 1954) given by:

$$\pi^*(s) = \underset{a}{\operatorname{argmax}}\, Q^*(s,a) \tag{3}$$

A PRC can be defined as a policy that maximizes the expected cumulative reward an agent can achieve for the given MDP. In other words, no other choice can provide a more expected cumulative reward for all defined state-action pairs of the given MDP. Later, more details on how to solve the MDP problem and find the optimal policy are discussed.

2.2. MDP and Reinforcement Learning Distinctions

Before moving to MDP in driving contexts, it is useful to discuss the distinctions between MDP and reinforcement learning (RL). In an MDP framework, shown in Figure 1.A, given (not learned nor estimated) states, actions, rewards and transition probabilities are used to obtain the optimal policy under each state. In RL however, as shown in Figure 1.B, either reward or transition probabilities or both are unknown (actions and states are given). Therefore, the agent learns the reward and transition probability through acting and experiencing i.e. exploration. That said, a typical RL algorithm comprises of two steps. The first step is to learn the missing MDP model elements, which are the reward and/or the transition probabilities. Second, the algorithm must solve the learned MDP to obtain the optimal policy given the states. Figure 1.B depicts the framework of an RL where the learning comes from a combination of exploration and exploitation. Learning solely through exploration might not maximize the reward because the agent tries to obtain more information about the reward by trying out things with unknown rewards. On the other hand, pure exploitation (making the best actions given the state obtained from the previously solved MDP) might lead to being stuck with policies with small amounts of reward. Typically, trade-offs between exploration and exploitation are used in order to benefit from both.

Figure 1.C presents the approach used in this paper. Using instrumented vehicle data, all of the inputs for the MDP framework are estimated (learned) first and then used to find the optimal policy. This approach is between MDP and RL (in a pure MDP, the rewards and transition probabilities are often given or pre-determined by researchers); after learning the reward and transition probability from observed data (similar to RL), all inputs become available in order to obtain the optimal policy (which is similar to MDP). Since the process of estimating MDP parameters, such as the rewards and transition probabilities, are pre-estimated from observed data, RL "learning" methods such as epsilon-greedy Q learning are not addressed in this paper. However, the obtained reward is different from the rewards of typical RL practices. The reward in RL is unknown to only the agent (known to the trainer, researcher, designer etc.) and once the agent acts and lands in a specific state, it realizes the reward associated with landed state and contributes that knowledge for the next decision. The reward in this paper, however, is unknown to both the driver (agent) and researcher and the drivers do not receive feedback in terms of actions that maximize their rewards. We estimated the rewards using a discrete choice model. Therefore, individual drivers have different rewards. Ideally, in order to train drivers with the purpose of making better driving decisions, or to train autonomous vehicle computers to behave similar to human drivers or to make safe decisions, RL is preferred because the rewards can be given to the agent (driver, computer) in real-time (Kamrani et al., 2018b) in terms of distance to the front objects, some calculated driving volatility (Kamrani, 2018, Arvin et al., 2019b, Kamrani et al., 2018a) and risk

factors (Arvin et al., 2019a, Kamrani et al., 2019) etc.

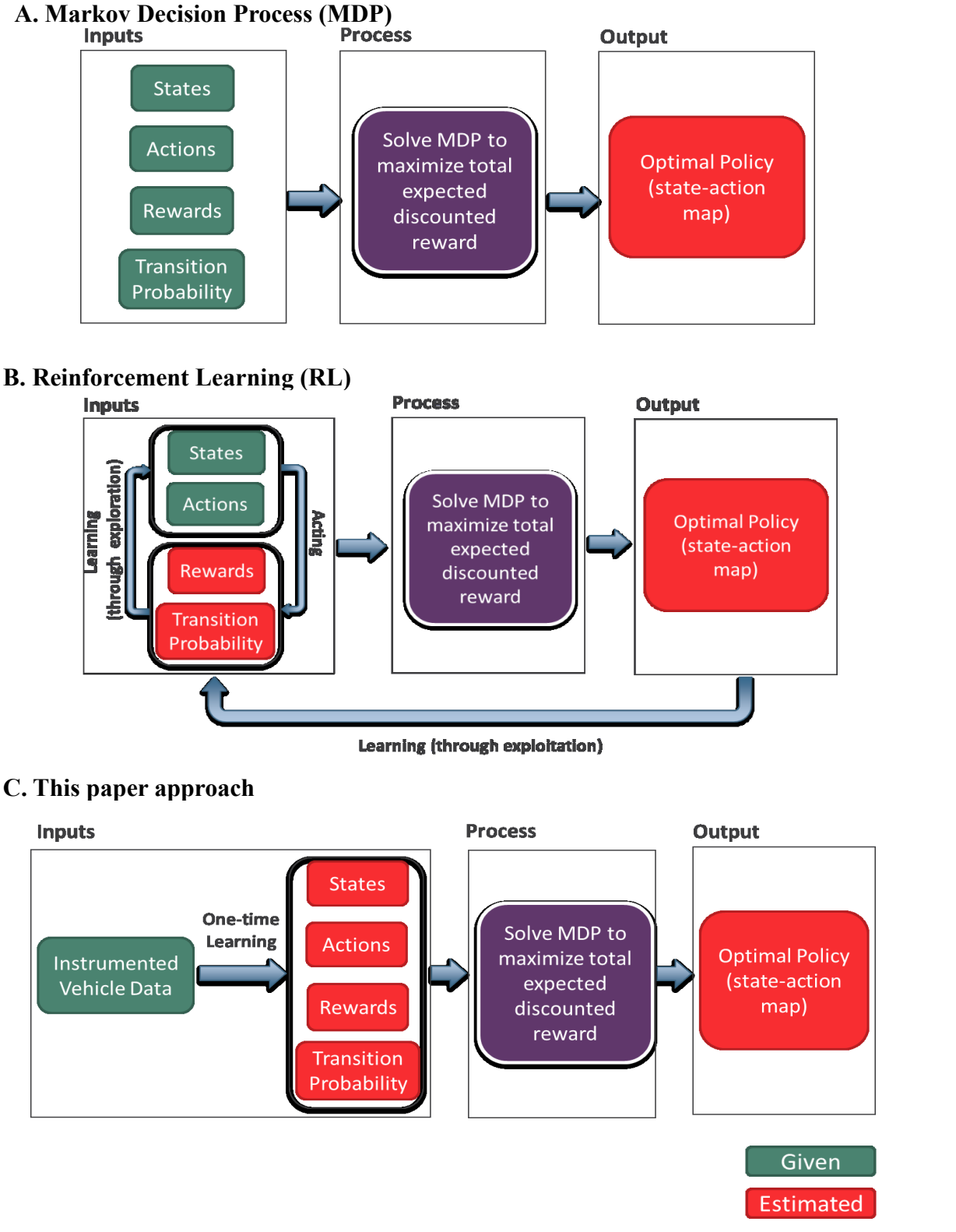


Figure 1. Framework of Markov Decision Process (A), Reinforcement Learning (B) and the Paper (C)

2.3. MDP in Driving Maneuvers Context

The MDP comprises of 1) the driver as the decision maker or agent; 2) states, which can be defined by surrounding traffic conditions, current vehicle speeds, and traffic along the route; and 3) a set of actions such as acceleration, deceleration, maintaining speed, and lane change. Depending on the defined states and actions, MDP can provide a transition probability matrix and a reward function which provide the numerical measures of payoff associated with transitions from one state to another. The state transition probability can depend on (but is not limited to) vehicle dynamics, current speeds of surrounding vehicles, minimum stopping distance, maximum acceleration or deceleration rates, distance between vehicles, reaction time, and time to collision.

2.3.1. Actions

The variability of drivers' actions can be attributed to several factors such as aggressiveness, gap acceptance, psychological states, mindsets, attitudes, and preferences. A preferable MDP set of actions for the explained setting include accelerating, braking, changing lanes to the left, changing lanes to the right, and maintaining constant speed. However, lane changing is excluded in this study due to the variable limitations in the available dataset. Therefore, the actions are:

- *A*: Acceleration; increasing speed at a rate higher than a specified threshold
- *C*: Maintaining Constant speed; no acceleration or deceleration, while maintaining speed within a specified threshold
- *D*: Deceleration; decreasing speed at a rate higher than a specific threshold

It is necessary to define thresholds in high-resolution data in order to overcome subtle deviations in speed that may be noise and hard to attribute to driver decisions, especially when acceleration values define driver decisions. As an example, let us say that a driver has activated cruise control or decided to maintain constant speed by not changing the acceleration pedal's displacement. If the magnitude of the acceleration during that time is recorded ten times per second, zero acceleration values will rarely be observed. To remove noise, the data is aggregated over one second and a threshold is used to characterize drivers' decision to maintain constant speed. Values of accelerations that fall between -2.5% to 2.5% of the data around zero are considered constant speed. Accordingly, positive and negative values of acceleration are defined as acceleration and deceleration decisions, respectively.

2.3.2. States

Figure 2 (left) illustrates the states of instantaneous driving decisions. The red car (host) is in the state (s) where there are four vehicles around it. The driver decides to change lanes and overtake the vehicle in front of it (the yellow car). Among all different outcomes (stochasticity), by choosing to change lanes he/she ends up in a new state, denoted by s' (which also depends on the other drivers' maneuvers and decisions). It should be noted that in a given state, different drivers make different decisions (i.e. take different actions). In this example, the driver of the red car decided to change lanes, but even maintaining constant speed can change someone's state due to the dynamics of surrounding traffic.

To show the states formally, a spatial layout of the proximate vehicles is illustrated in Figure 2 (right). The central cell of the matrix represents the host vehicle. The surrounding cells represent the surrounding space, which can be empty or occupied by a vehicle or other objects (0 for empty and 1 for occupied). There are eight cells around the host vehicle; each can either be 0 or 1 forming $2^8 = 256$ different possible states. The number of possible states will decline if a vehicle is in the left or right lanes or if there are only two lanes in the same direction (32 states).

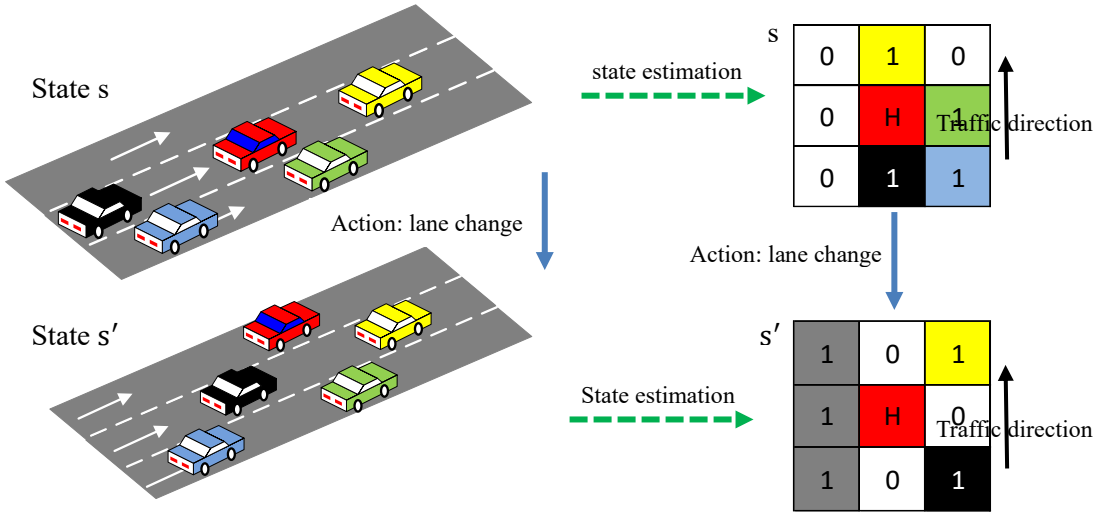
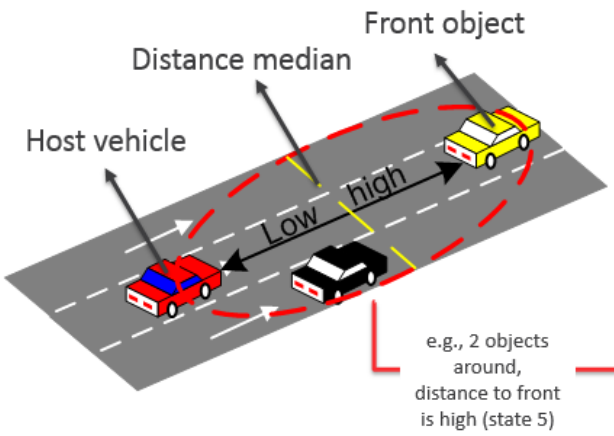


Figure 2. Examples of the Driving States and Actions taken by Drivers

Determining the current state (one of the 256 possibilities) of a vehicle, based on the structure shown in Figure 2 requires a complete awareness of its surroundings (360 degrees). The technology that provides such awareness and related data is already available in certain instrumented vehicles (IVs), which can create vision-based high-density maps through cameras and laser scanners. The data at hand only provides the total number of objects around the host vehicle without their spatial positions and distance to the front object. Due to one-second data aggregation, the average number of objects and the average distance to the front object (over one second) define the states (Figure 3).

Figure 3 also provides an example of how to determine the host vehicle's state. There are two objects around the host vehicle; therefore, according to the table on the right, rows 5 and 6, which correspond to the number of objects between two and three, should be referenced. Since the front vehicle (yellow car) is beyond the distance median, the current state of the host vehicle is determined as five.



State	Distance to front object	Ave. No. of objects (during 1 sec.)
1	High	[0 , 1)
2	Low	[0 , 1)
3	High	[1 , 2)
4	Low	[1 , 2)
5	High	[2 , 3)
6	Low	[2 , 3)
7	High	[3 , …)
8	Low	[3 , …)

Increased level of crowdedness

***Note**: Low and high in the table are different for each trip.*
***High**: Distance to the front object ≥ median distance*
***Low**: Distance to the front object < than median distance*

Figure 3. Definition of States Based on Number of Objects and Distance to the Front Object

2.3.3. Transition Probabilities

Figure 4 depicts an example of the MDP structure where the states are shown by circles, decisions by dashed arrows (A: acceleration, D: deceleration, C: constant speed) and transition probabilities by solid arrows. For each set of initial states, actions and landed states (s, a, s'), a transition probability is defined as $T(s, a, s')$. This quantifies the probability of ending up in state s' given that the agent's initial state was s and it took action a i.e. $P(s'|s, a)$. For each trip, having determined the states (according to Figure 3) and actions (according to acceleration values), the probabilities are specified by counting the occurrence of each set (s, a, s') and dividing them by total number of possible occurrences. Therefore, there are three matrices (for acceleration, deceleration and maintaining constant speed) each 8 by 8, where rows are the initial states (1 to 8) and columns are the landed states (1 to 8). For example, the element (1, 2) in acceleration matrix indicates the probability of transitioning from state 1 to state 2 (the driver accelerated).

2.3.4. Reward

Consider an aggressive driver who tends to reduce speed later than a calm driver, or may drive in close proximity to other drivers. Such behaviors are likely to be personally optimal for him/her. Therefore, the reward in a driving context is unknown, and varies from one driver to another. Even for the same driver, it might change depending on several factors such as trip purpose, traffic condition, time of the trip, etc. The goal is to find a reward structure based on the personal reward function that uncovers PRCs, which drivers use to make instantaneous driving decisions. In the MDP structure of this study, the reward function is dependent on both the taken action and the landed state.

***Note**: To avoid clutter, transitions from State 1, given acceleration decision and related rewards are shown.*
***A**: acceleration*
***D**: deceleration*
***C**: constant speed*
***Circles** (1 to 8): states*
***T(s, A, s')**: probability of transition from sate s to s' given acceleration decision*
***R(A, s')**: reward of landing in state s' given acceleration decision*

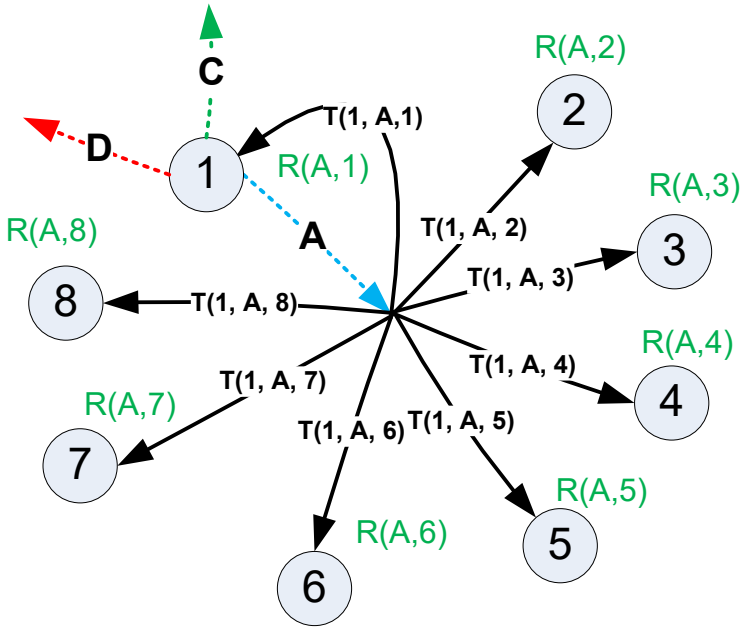


Figure 4. MDP structure used in the study

Utility functions have long been used in transportation to represent and explain travel behavior (e.g. mode choice) (Train, 1986, Train and Winston, 2007, Train, 1978, Ramming, 2001). In such cases, there are clear economic arguments for specifying a utility function. For example, a choice between using a bus vs. car for daily commutes is likely made by considering the respective cost and time, the comfort, and convenience of the modes for specific trip purposes. Multinomial logit (MNL) model is justified for reward estimation, given that the model provides a framework for calculating relevant probabilities using a simple mathematical form. This framework can be further expanded to include relevant variables that associate with reward probabilities. To estimate the reward for each state-action pair, the MNL model is applied separately for each trip. In other words, 120 MNL models are estimated by treating each second of each individual trip as one observation. In this setting, observations are assumed independent from each other. The dependent variable (discrete choice) is the driver's action and the current state is considered as independent categorical variable. The utility perceived by the driver consists of an observable component and an unknown random error. Therefore, the utility can be represented as (Koppelman and Bhat, 2006):

$$U_{it} = V_{it} + \varepsilon_{it} \tag{4}$$

where U_{it} is the utility of the alternative i for driver at time t,
V_{it} is the observable portion (systematic component) of the utility, and
ε_{it} is the error or unknown component of the utility.

The systematic component of the utility can be expressed as:

$$V_{it} = V(X_i) + V(S_d) \tag{5}$$

where V_{it} is defined as before and $V(X_i)$ is the portion of utility of alternative i associated with the attributes (or states) of alternative i. Also $V(S_d)$ represents socio-economics of the driver d, though

such information was not available in the data analyzed. The importance of states can be estimated as:

$$V(X_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \quad (6)$$

where β_k is the parameter which estimates the strength of association of state k on the utility of an alternative and X_k is the presence or absence of state k (there are 8 states based on the number of surrounding objects and distance to the front object) for alternative i. In this context, the utilities of the alternatives to Accelerate (A), Decelerate (D), or maintain constant speed (C) are shown as follows with alternative C serving as the base:

$$V(X_A) = \beta_{0,A} + \beta_{1,A} State\ 1 + \cdots + \beta_{7,A} State\ 7 \quad (7)$$
$$V(X_D) = \beta_{0,D} + \beta_{1,D} State\ 1 + \cdots + \beta_{7,D} State\ 7 \quad (8)$$
$$V(X_C) = 0 \times State\ 1 + \cdots + 0 \times State\ 7 \quad (9)$$

where $\beta_{0,A}$ and $\beta_{0,D}$ are the estimated intercepts for acceleration and deceleration alternatives, respectively; $\beta_{*,j}$ are parameter estimates for state-specific dummies (for 8 states, we have * = 1 to 7 dummy variables, j = acceleration, deceleration, and maintaining constant speed). Note that State is a categorical variable and therefore we use n-1 categories in the MNL model, with least crowded and long distance to the front object as the base.

If the utility of acceleration is highest among the alternatives, then a driver chooses acceleration. The basic equation defining an MNL model probability of choice between alternatives is given by:

$$P_i = \frac{\exp\{V_i\}}{\sum_{j=A,D,C} \exp\{V_j\}} \quad (10)$$

where P_i is the probability of choosing i among j choices and V is the systematic component of the utility (the error component is Gumbel distributed). This equation (10) indicates that if an alternative has a higher utility, its probability of being chosen is higher. That said, by estimating the coefficients of the MNL model, the probability of driver actions given the state are used as the reward in the MDP model:

$$R_i = \frac{\exp\{V_i\}}{\sum_{j=A,D,C} \exp\{V_j\}} \quad (11)$$

Using the probabilities obtained from the MNL model as respective rewards assumes that the reward of an action given the state is proportional to the probability of that action given the state. Notably, several assumptions of the logit model such as the Independence from Irrelevant Alternatives can be investigated further by using nested structures. And the panel nature of the data can be addressed by estimating panel-data mixed logit models. Given the focus of the study on

demonstrating how the rewards can be linked with microscopic behavioral decisions, the simple MNL model was deemed sufficient. Now, we have all the necessary inputs to solve the MDP problem, i.e. states, actions, transition probability and a reward function. Next, we discuss the solution approach called value iteration.

2.4. Value Iteration

In the context of this study, solving an MDP is equivalent to finding the PRCs. The value iteration algorithm (Bellman, 1954) is used to find the PRC for each state. The steps of the algorithm are as follows:

Step 1: For each state ($s \in S$) initialize $V(s) = 0$
Step 2: Set a threshold (θ) as stopping condition
Step 3: For each state ($s \in S$) – (loop over states):
3.1 $\Delta \leftarrow 0$
Repeat – (loop over actions):
3.2 $v \leftarrow V(s)$
3.3 $V(s) \leftarrow max \sum_{s'} T(s, \pi(s), s')[R(\pi(s), s') + \gamma V(s')]$
3.4 $\Delta \leftarrow max\ (\Delta, |v - V(s)|)$
Until $\Delta < \theta$
Step 4: PRC $= \pi^*(s) = \underset{a}{argmax} \sum_{s'} T(s, \pi(s), s')[R(\pi(s), s') + \gamma V(s')]$

Given each state, the algorithm gives one of the actions (either acceleration, deceleration or maintaining constant speed) as the optimal policy. As discussed earlier, since the reward function estimates are different from one driver to another, we call the optimal policy as the Personally Revealed Choice (PRC) as opposed to the optimal policy.

A toy example was used to verify the performance of the algorithm. Specifically, Figure 5 depicts a grid world as a classic example in MDP, where an agent can either go up, right or stay. The cells of the grid are different states. The agent's action outcome is stochastic, which is defined based on the transition probability matrix shown in Figure 6. If the agent decides to go up, there is a 70% chance it ends up in the upper cell and has equal chances of 10% of either staying in its current state, moving left, or moving right. Similarly, if it decides to go right, 70% of the time it ends up in the right cell and there are equal chances of 10% that it ends up in its current cell, or above or below. In the case where the agent decides to stay, it will always remain in its current cell. Certain outcomes are impossible in some states. For example, if the agent is in state two and decides to " go right," given that the agent's right and bottom sides are blocked, the probability of landing in the right cell (70%) and landing below (10%) will be distributed to other available outcomes equally (in this example, the probability of landing in 4 or staying in 2 become 50% each, i.e. 10+(70+10)/2. The rewards for states (shown in parentheses) are zero except for states five and eight where their rewards are -1 and +1 respectively. This reward structure implies that the optimal policy is the one that moves the agent toward state eight to maximize the cumulative discounted reward. Therefore, the optimal policy (shown in arrows) for the agent is to go up when

in states 1 to 5, go right in states 6 and 7, and remain when in state 8. This toy example was introduced to the algorithm to assure that the algorithm also produces optimal policies, shown as arrows in Figure 5. By changing the rewards, we also checked if the algorithm yields respective intuitive optimal policies.

3. 6 (0) 4. →	7 (0) →	8 (+1) stay
3 (0) ↑	4 (0) ↑	5 (-1) ↑
1 (0) ↑	2 (0) ↑	

(**Note**: states from 1 to 8, reward are shown in parentheses, arrows represent optimal policies)

Figure 5. Grid World Toy Example

Go up		Landed state (s')							
		1	2	3	4	5	6	7	8
Initial state (s)	1	0.15	0.15	0.7	0	0	0	0	0
	2	0.15	0.15	0	0.7	0	0	0	0
	3	0	0	0.15	0.15	0	0.7	0	0
	4	0	0	0.1	0.1	0.1	0	0.7	0
	5	0	0	0	0.15	0.15	0	0	0.7
	6	0	0	0	0	0	0.5	0.5	0
	7	0	0	0	0	0	0.33	0.34	0.33
	8	0	0	0	0	0	0	0.5	0.5

Go right		Landed state (s')							
		1	2	3	4	5	6	7	8
Initial state (s)	1	0.15	0.7	0.15	0	0	0	0	0
	2	0	0.5	0	0.5	0	0	0	0
	3	0.1	0	0.1	0.7	0	0.1	0	0
	4	0	0.1	0	0.1	0.7	0	0.1	0
	5	0	0	0	0	0.5	0	0	0.5
	6	0	0	0.15	0	0	0.15	0.7	0
	7	0	0	0	0.15	0	0	0.15	0.7
	8	0	0	0	0	0.5	0	0	0.5

Figure 6. Transition Probabilities of the Toy Example (the transition probability of action "stay" is an identity matrix)

5. Data

The data used in this study relates to the Basic Safety Messages (see Figure 7) sent and received by vehicles that participated in the Safety Pilot Model Deployment (SPMD) in Ann Arbor, Michigan (Henclewood, 2014). The data is stored in the Department of Transportation ITS JPO (Joint Program Office) Data website (https://data.transportation.gov/Automobiles/Safety-Pilot-Model-Deployment-Data), maintained by the U.S. Department of Transportation. The data was collected for two months (April and October 2012) in the real-world from more than 2,800 vehicles equipped with DSRC devices that transmitted instantaneous vehicle geocodes and kinematics such as speed, acceleration, heading, yaw rate, etc. This study uses a subset of SPMD data, collected by vehicles equipped with Data Acquisition Systems (DAS). It includes vehicle position (altitude, latitude, and longitude), motion (speed and acceleration), status of major components (accelerator, brakes, lights, cruise control, and wipers), and instantaneous driving contexts (surrounding objects and distance to the front object with an approximate detection range of 256 ft./78 meters). There were 259 trips undertaken by 71 unique vehicles. Since we are interested in drivers' decision-making driver mechanisms, the 10 Hz data (10 observations per

second) were aggregated to one observation per second. Aggregation also helps remove some noise in acceleration values, as discussed earlier. After exploring the data, 120 trips were found to be appropriate for this study because in order to obtain a reasonable transition probability, the driver has to encounter different states during his/her trip.

Figure 8 shows how the data needed for the analysis were prepared. Driver decisions are determined from the value of acceleration (step 1). In addition, each vehicle's current state is specified by examining the average number of objects, average distance to the front object, and by referring to Figure 3 (step 2). Landed state, which is the stochastic outcome of the driver's decision, is the one-second lag of the current state (step 3). Red arrows on Figure 8 show the process.

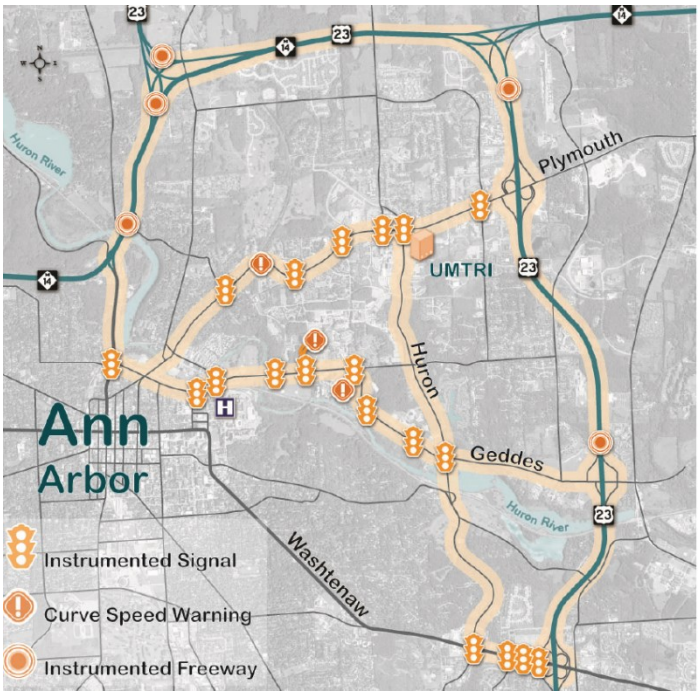


Figure 7. Safety Pilot Model Deployment Site, Ann Arbor, Michigan (Henclewood, 2014)

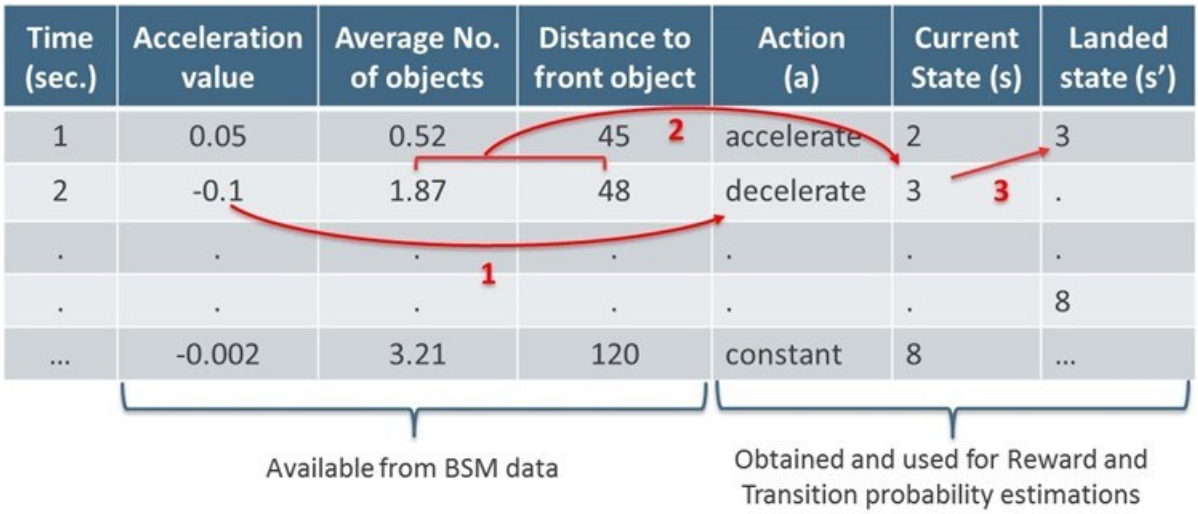


Figure 8. The data preparation steps

6. Results

6.1. Data Descriptive Statistics

Table 1 provides the descriptive statistics of the 120 processed trips. The results seem reasonable (e.g., average speed is about 40 mph and average trip duration is 18 minutes). On average, at the trip level, there is an object in front of the host vehicle 77% of the time. The distance to front object on average is 44.2 ft, with a standard deviation of 20.43 ft.

Table 1. Descriptive Statistics of the Selected Trips (n=120)

Variable	Explanation	Mean	SD	Min	Max
Speed (mph)	Average vehicle speed during the trip	40.49	35.12	3.10	70.61
Acceleration (ft/s^2)	Vehicle acceleration during the trip	1.05	0.62	0.26	4.46
Deceleration (ft/ s^2)	Vehicles deceleration during the trip	-1.48	0.95	-6.20	-0.29
Duration (min)	Duration of the trip	18.16	32.96	1.32	217.53
Average Number of Objects	Average number of objects around the host vehicle during the trip	1.63	0.88	0.03	3.53
Object present (yes)	Binary variable indicating if there is an object in front of the host vehicle	0.77	0.25	0.03	1
Distance to Front Object* (ft.)	Average distance from the head of host vehicles to the front object during the trip	44.20	20.43	11.38	94.57

**Only for cases when a front object is present.*

6.2. Discrete Choice Model

The MNL model is used to estimate rewards by assuming that rewards are proportional to the probability of actions given the state. For each of the 120 trips, an MNL model is estimated where the choice outcomes are acceleration, deceleration and maintaining constant speed (as the base), and the independent variable is the categorical variable of state. Table 2 presents the descriptive statistics of the 120 MNL estimates. The table indicates, on average, that acceleration decisions are less likely to be taken than maintaining constant speed decisions in states 2, 3 and 4 (compared

with state 1). In states 5 to 8, however, an acceleration decision is more probable than maintaining constant speed. Similar interpretations can be seen for deceleration decision using the mean of respective coefficients in Table 2. The results show a wide range of estimates with high standard deviations across trips encompassing positive and negative coefficients, making the results about overall driver preference inconclusive.

Table 2. Descriptive Statistics of MNL Estimated Coefficients (n=120)

Choice	Variable	Category	Estimate Mean	Estimate Std. Dev.	Min	Max
Constant Speed (base)		--	--	--	--	--
Acceleration	Intercept	--	0.97	3.86	-2.56	20.15
	State	1	--	--	--	--
		2	-0.68	4.8	-19.34	19.31
		3	-0.9	4.33	-18.77	14.53
		4	-0.39	5.05	-20.15	18.37
		5	0.15	5.53	-32.13	17.5
		6	0.58	7	-19.87	18.38
		7	0.44	6.1	-20.31	18.29
		8	0.23	7.07	-34.53	18.98
Deceleration	Intercept	--	0.21	4.48	-18.6	20.15
	State	1	--	--	--	--
		2	0.21	6.05	-20.15	36.68
		3	-0.43	5.22	-19.52	18.04
		4	-0.11	5.17	-19.25	18.93
		5	1.3	6.27	-19.72	19.73
		6	0.92	7.07	-20.15	19.61
		7	0.98	6.23	-19.81	20.15
		8	0.73	7.92	-35.22	18.98

Notes: State 1 is least crowded for the subject vehicle and State 8 is most crowded. The estimates presented are the results of 120 MNL models. The Min/Max values come from model(s) whose coefficients may not be statistically significant (5% level).

Although the estimated coefficients of state-action pairs for individual trips could be used as respective rewards, they all are zero for the base choice (constant speed) and they change if we consider acceleration or deceleration as the base. Therefore, the probability of choices given the states, which is invariant to the base choice selection, is used as per Equation 11. Figure 9 presents boxplots of estimated rewards of states 1 and 8 for maintaining constant speed and deceleration acceleration actions. The figure shows that the acceleration decision in both states 1 and 8 has a higher median compared to other decisions. However, when it comes to state 8 which is most crowded state in terms of traffic, the median of constant speed reward (0.09) is considerably lower than the acceleration reward (0.31) and the deceleration reward (0.26). This is reasonable because, in heavier traffic, drivers stop and go more frequently than travel at constant speed. Therefore, in state 8, we would expect a lower reward (i.e. probability) for traveling at constant speed.

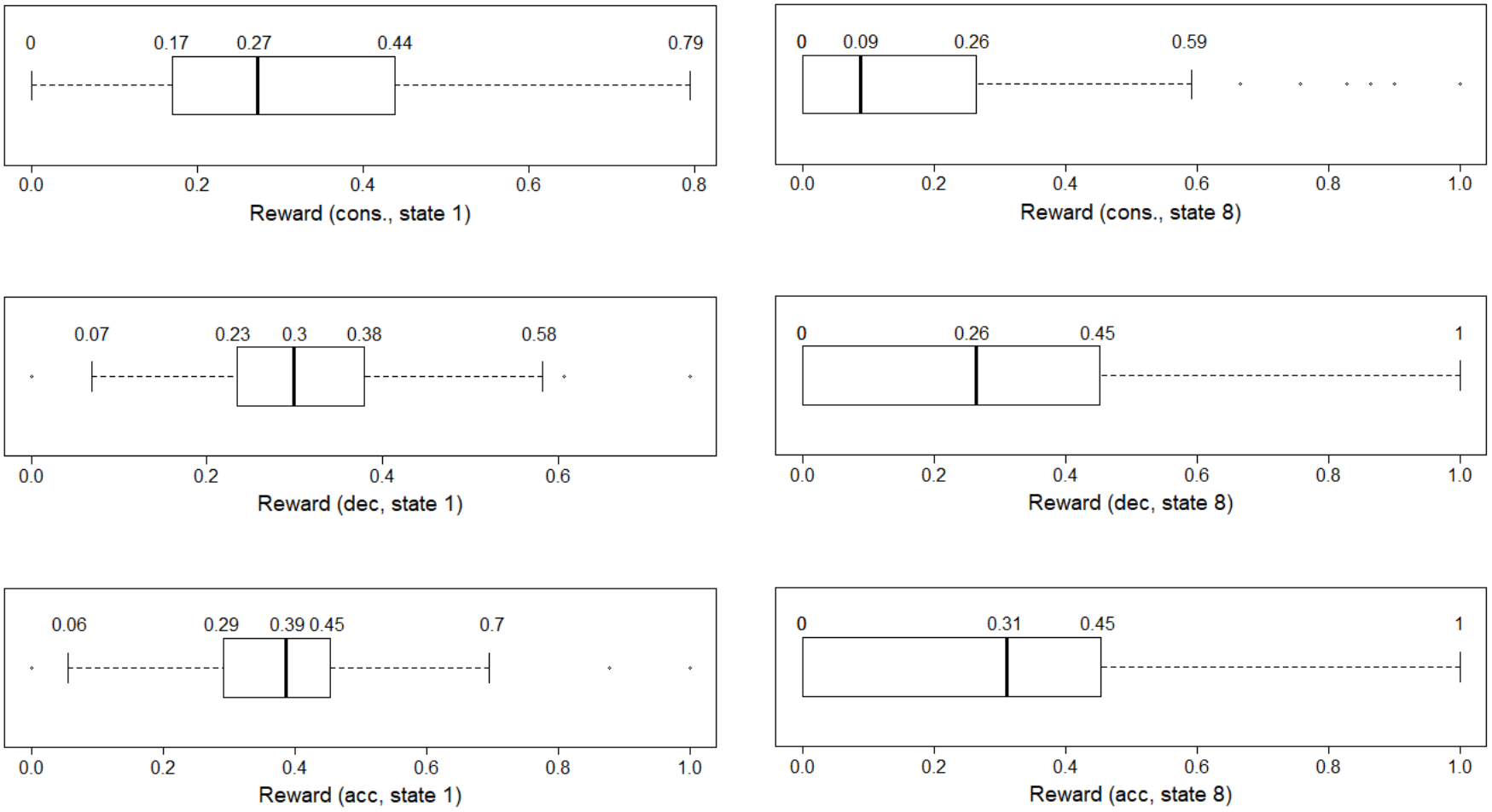


Figure 9. Reward Distribution of States 1 and 8 for maintaining Constant Speed, Acceleration and Deceleration (n = 12)

Having estimated the rewards, the question is what action given state maximizes the total discounted reward. Note that although drivers make decisions, they do not have control of the landing states because of the stochastic nature of driving environments and decision outcomes. Using the state-action transition probability coupled with state-action rewards, we obtain the optimal policy (i.e. PRC) for individual drivers.

6.3. Personally Revealed Choices (PRCs)

The estimated rewards along with the other inputs are introduced to the value iteration algorithm in order to obtain the optimum policies that maximize the total expected accumulated discounted reward. A discount factor of 0.95 was used in this study. The discount factor is an important parameter since it determines the relative weights of immediate versus long term planning. For each trip, the value iteration algorithm yields a PRC given the state. There is no predetermined goodness for each state. This means a driver might be OK with other vehicles around while another driver would rather avoid them. Some of these personal preferences are captured through the reward values.

Importantly, as we assumed that the reward is proportional to the estimated MNL probabilities, one possible way of maximizing the total accumulated reward is to select the actions with the highest probabilities (i.e. highest reward) as the PRCs. With this naïve approach, however, the stochasticity of decision outcomes, the sequence of decisions, and the future expected rewards are not accounted for. For comparison purposes, we have provided the results based on this approach in Figure 10, left.

Figure 10, right, presents the PRC results from MDP. We have shown the proportion of PRCs, which add up to 100% for each state. Figure 10's caption explains how to read the figure. The peaks and valleys in both figures are intuitive because the average number of objects are the same for states 1 and 2 and the only difference is the distance to the front object i.e. in state 2 the front object is closer to the host vehicle. This holds true for state pairs of (3, 4), (5, 6) and (7, 8). That naïve approach of PRC determination results does not show much trend in the proportion of PRCs. However, the MDP results show an ascending trend in the proportion of acceleration. Overall, acceleration is the most probable PRC throughout the states (ranges from 35% in state 2 to 58% in state 7) while the proportion of deceleration does not change significantly across the states.

Drivers have different levels of calmness and aggressiveness. An aggressive driver's PRC could be acceleration even when they are surrounded by three or more vehicles (i.e. state 7 or 8). Moreover, the layout of objects around the host vehicle also affects PRCs. If the objects were mostly in front of the vehicle throughout a trip, an intuitive PRC would suggest maintaining constant speed or deceleration. However, if the objects are on the rear or sides of the host vehicle, then perhaps acceleration might be the PRC in order to avoid a crash and crowdedness. That said, even the same layout of objects for two different drivers can result in different PRCs due to their differences in perceptions, information processing, situational factors, preferences, and

experiences. Given all the complex factors contributing to PRC results, the overall outcomes are reasonable and intuitive.

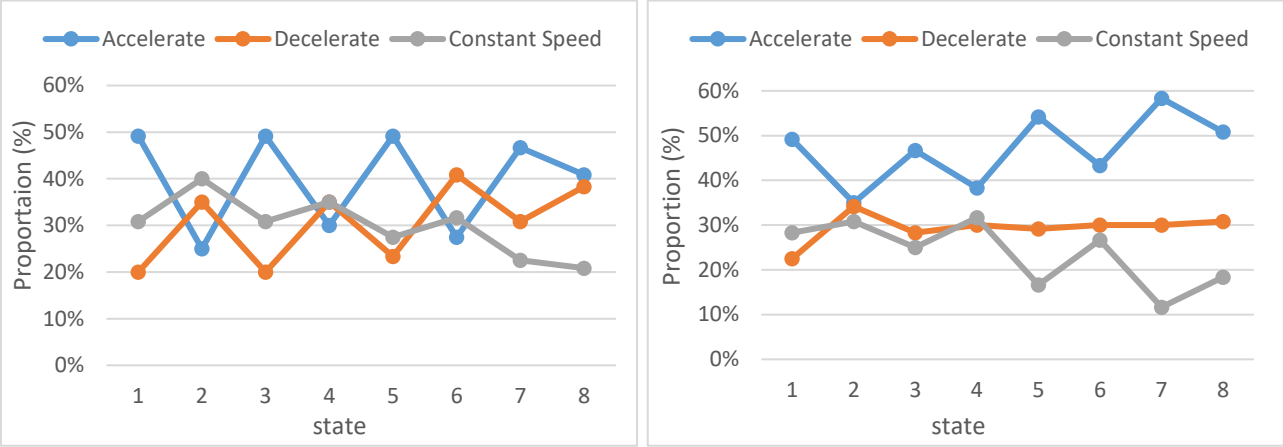


Figure 10. Personally Revealed Choices given specific states (left: naïve approach, right: MDP) (n=120 trips). Example of reading the figure (right): in state one, 49% of drivers' PRC was acceleration, 23% was deceleration and 28% maintaining constant speed.

6.4. PRCs at Different Levels of Crowdedness

To supplement information on PRCs with respect to changes in the number of objects, trips were segmented based on the level of crowdedness. On a crowded roadway where the average number of objects around a vehicle throughout a trip is higher, the distribution of PRCs may be different than a trip taken during less crowded conditions. Figure 11 presents the distribution of the average number of objects. Based on the distribution, three levels of crowdedness are defined: non-crowded, semi-crowded and crowded. Non-crowded trips have an average number of objects between 0 and 1, semi-crowded trips have an average number of objects between 1 and 2, and crowded trips have an average number of objects of more than 2.

In non-crowded trips (Figure 12), an ascending trend of the acceleration proportion from states 2 to 8 is obvious. This observation is very interesting because as the surroundings of the host vehicle become more crowded, the majority of drivers' PRC is acceleration. The intuitive interpretation of this behavior is that they prefer to avoid (or escape from) crowdedness by increasing their speeds. The successfulness of that decision depends on many factors such as instantaneous traffic conditions, surrounding drivers' decisions, etc. Although a PRC may not yield a drivers' desired outcome, the driver perceives the highest personal satisfaction from that decision compared to the other available options.

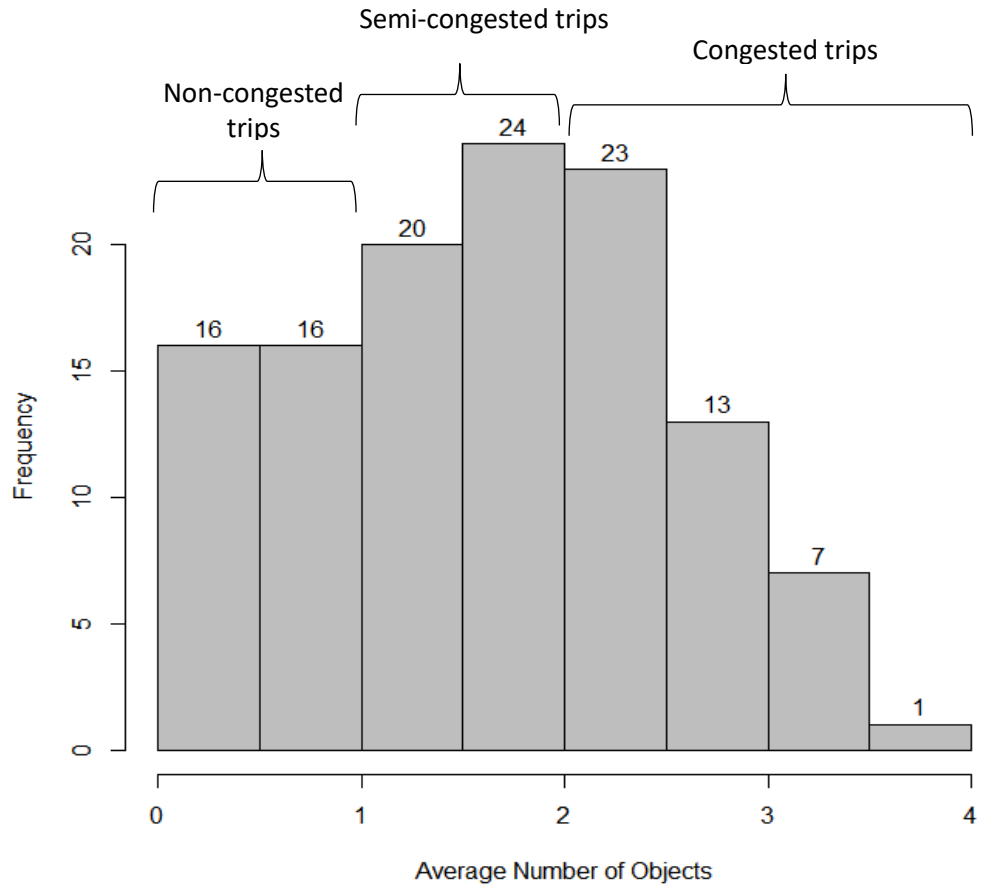


Figure 11. Distribution of Average Number of Objects Surrounding the Host Vehicle (n=120 trips)

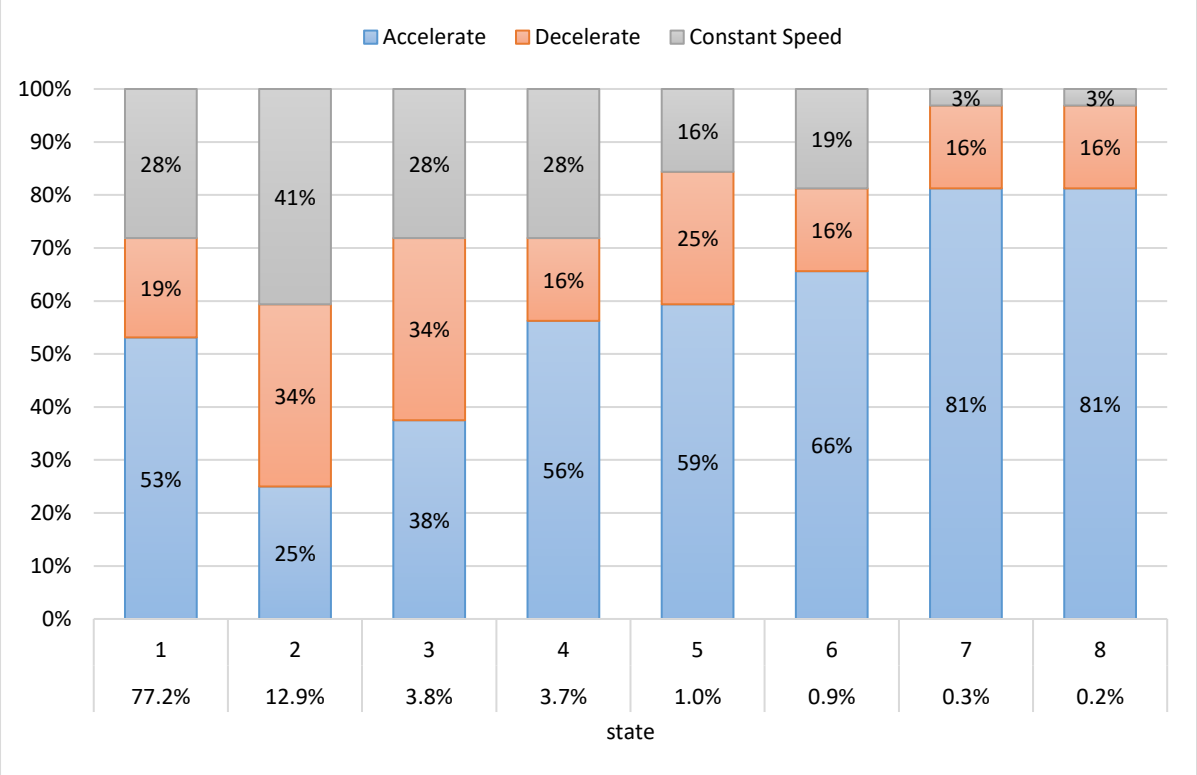


Figure 12. Personally Revealed Choices made on non-crowded Trips (n=32 trips)

In the semi-crowded trips results (Figure 13), state 4 has a balance among the proportions of each decision. After that, the proportion of the deceleration decision remains almost the same but around 10% from the constant speed proportion is replaced with the acceleration decision as the level of crowdedness around the host vehicle increases. Comparing non-crowded with semi-crowded trips shows that the proportion of acceleration has almost shrunk to half, indicating that more drivers considered deceleration or constant speed as their PRCs.

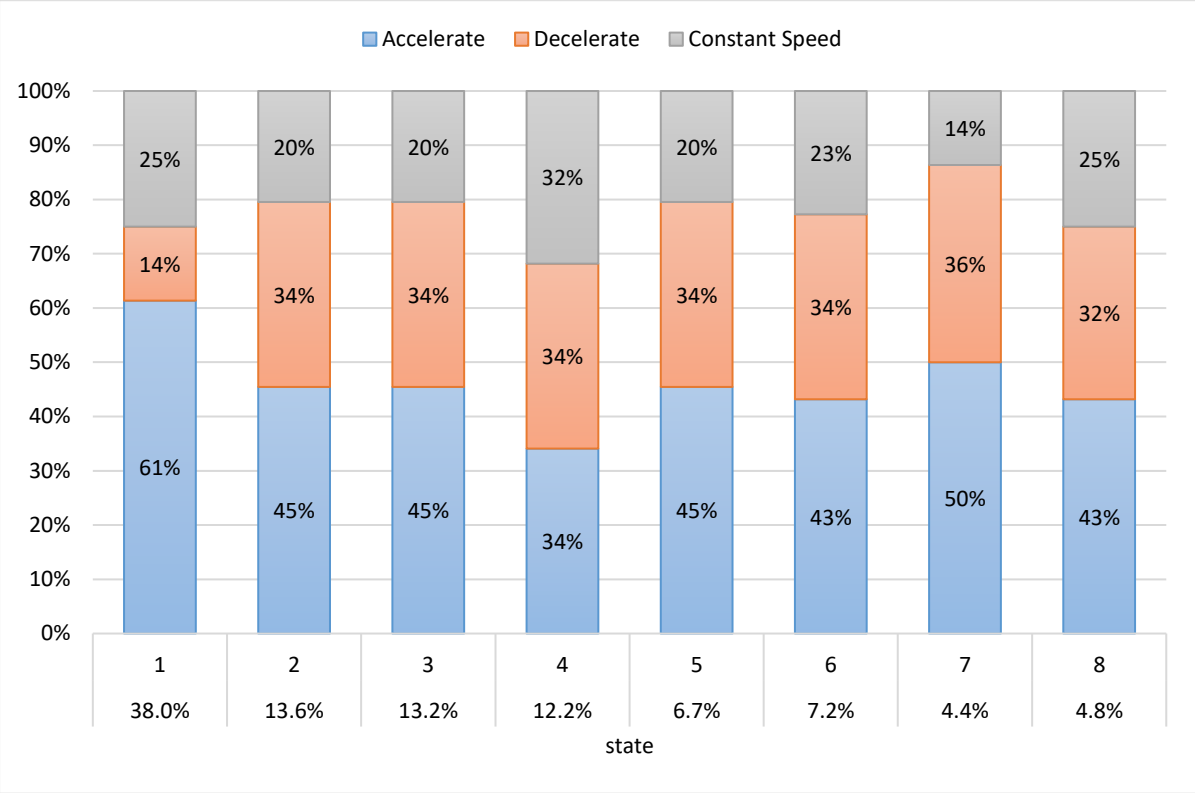


Figure 13. Personally Revealed Choices made on semi-crowded Trips (n=44 trips)

In crowded trips (Figure 14), the effect of trip crowdedness is more distinguishable. The acceleration decision is only dominant in states where the front object is far from the host vehicle. In our state definitions, each state pair [(1, 2), (3, 4), (5, 6) and (7, 8)] has a similar number of objects but the front object is closer to the host vehicle in the higher state. The impact of the front object can easily be seen by paying attention to the oscillations of acceleration deceleration and constant speed proportions across the states. In other words, closer front objects increased the rewards for deceleration and constant speed for some drivers, which is intuitive. This change in the reward can be due to two reasons. First, some drivers intentionally decide to replace acceleration with deceleration and constant speed because of the more crowded traffic conditions. Second, even though some drivers have not chosen to do so, the traffic conditions impose the above-mentioned replacement on them. In other words, drivers may be forced to follow (or be more constrained) rather than having the option to make decisions freely.

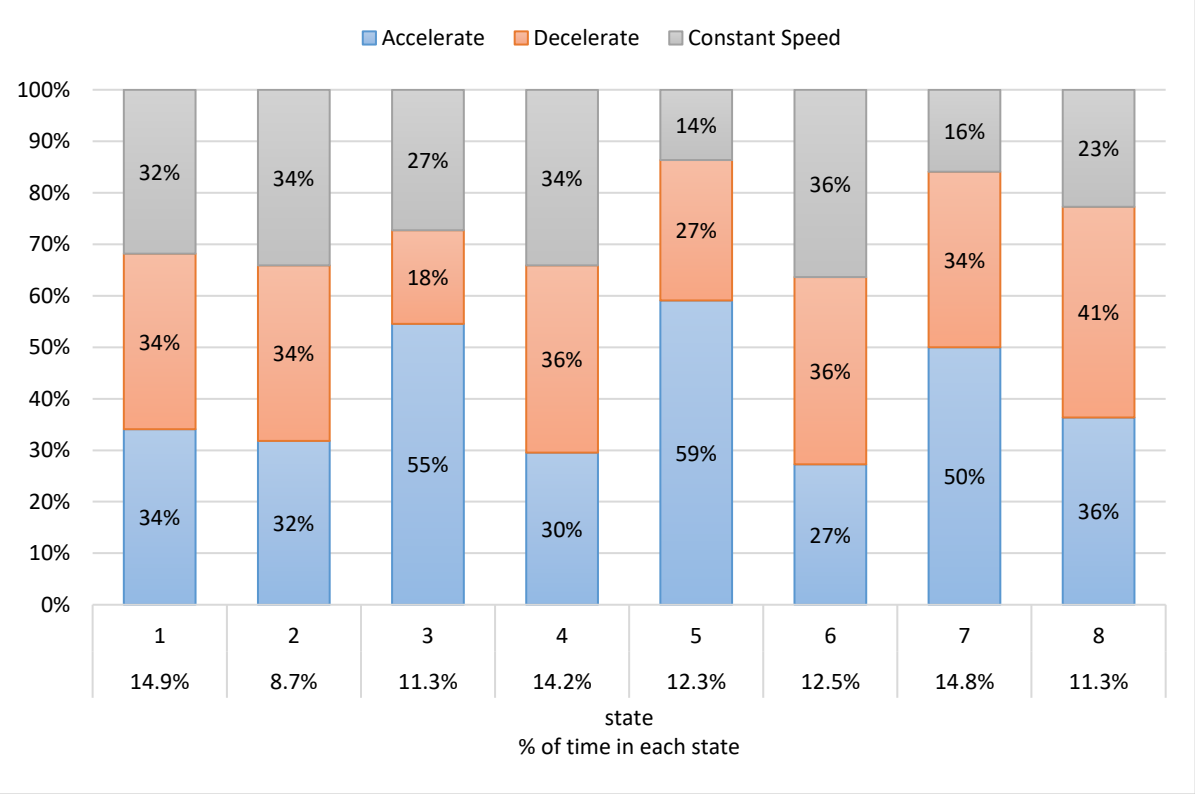


Figure 14. Personally Revealed Choices in crowded Trips (n=44 trips)

6.5. Comparison of PRCs and Real-world Observed Choices

Having found PRCs for each trip, it is beneficial to observe how long each driver can follow his/her PRC. It is evident that even though PRCs are obtained through the MDP method, drivers are not able to follow their PRCs all the time. For instance, a driver could prefer to accelerate in a particular state and that decision would be preferred, but traffic conditions (and many other hidden reasons) make the driver constrain the driver and compromise his/her PRC at that moment i.e. the driver cannot behave optimally. In fact, investigating the level of PRC following can help us see to what extent a driver's behavior matches their PRCs, i.e. how much of observed behavior are explained by PRCs.

Figure 15 presents the distribution of PRC following among the 120 trips. The numbers above the bars indicate the number of trips and each bar label on the horizontal axis shows the percentage range of time where drivers followed their PRCs. For example, the bar shown in the red box indicates that drivers could follow their PRCs, on average, about 70% of the time (the bar midpoint) in two of the trips (the range is between 66% and 74%). Likewise, if we cumulatively look at the top three bars in Figure 15, the level of PRC following is at least 30% in 71% of the trips (85 trips). The weighted average (based on trip duration) of following PRCs is 36%, meaning that the proposed PRCs are followed across the trips 36% of the time. Before we discuss the potential application this study in the conclusion section, we will discuss its limitations. Given that the study is based on applying a theoretical model and using detailed data, reasonableness checks, and the intuitive consistency of the results and interpretation are the primary validation tests. The

results of this study are consistently intuitive, and the results are reasonable and consistent with the findings of previous studies.

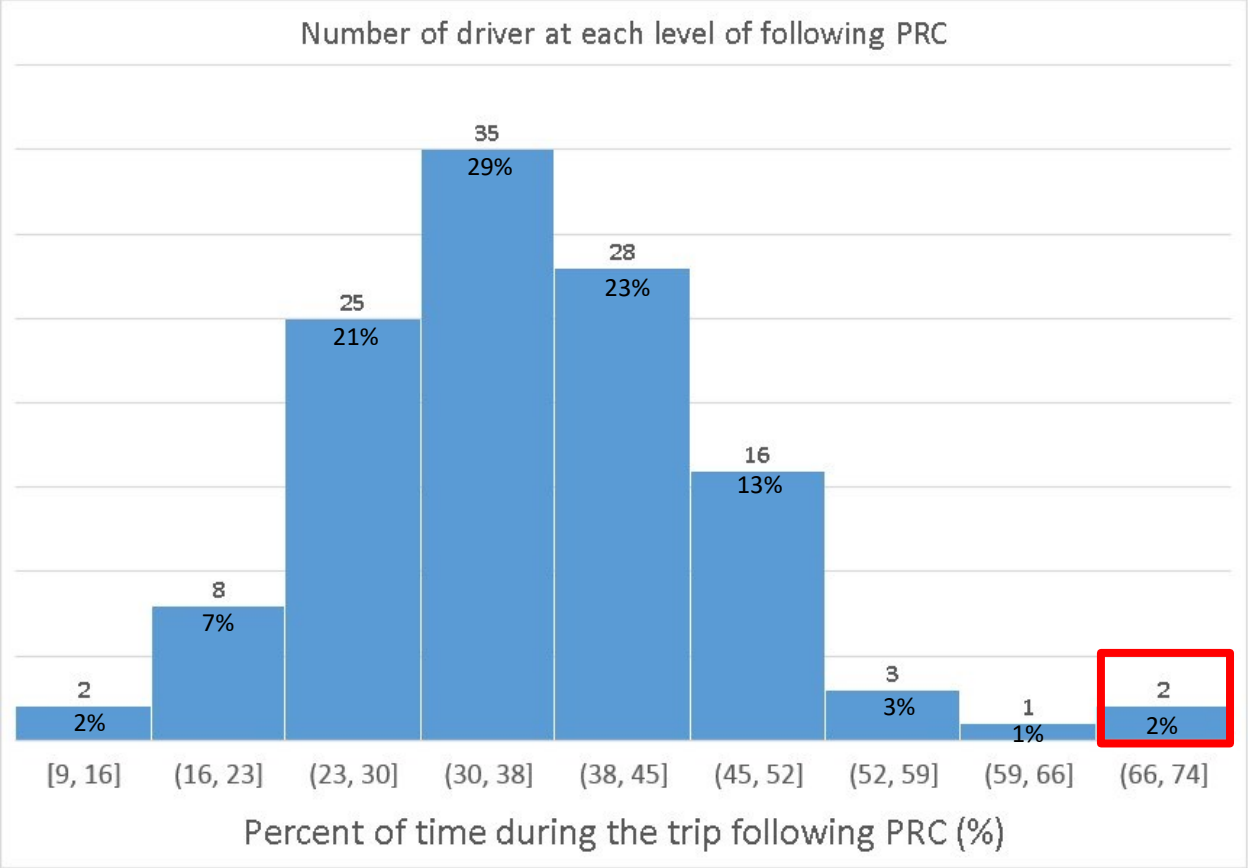


Figure 15. Distribution of following PRC among Drivers (n=120)

7. Limitations

The significant limitation of this study lies in the data. Although the DAS dataset provides relatively accurate data for acceleration and speed, the amount of information about the objects surrounding the host vehicle is limited. Only the number of objects and distance to the front object are provided. As was noted in the methodology section, the authors had to deviate from the complete method for defining the states due to the lack of information on object positions around the host vehicle. Notably, the state vector could be expanded with additional data (besides distances to the front object and number of surrounding objects) to obtain a more accurate representation of the state of the driver. Having a simpler model to partially explain driving behavior or the environment has plenty of precedence e.g., a lot of car-following models that are being used today are simple representations of real-world conditions and ignore many factors contributing to a host vehicle's acceleration and deceleration decisions. Thus, we recognize that the number of states (ideally) used to represent the problem are prohibitively large, but with simplifications we have reduced the number of states to a manageable size.

The data limitation also constrained the possibility of defining more states and, consequently,

more sophisticated reward functions. For the purpose of this study, the reward function was assumed to be dependent on the decision made and the landed state in order to simplify the computation. In addition, the lack of driver perception-reaction times in the data and consequently in the models are recognized as another limitation of this study. It is also acknowledged that the data in hand does not have any information on factors that impact the transition from one state to another leading to a method to obtain transition probabilities that does not captures all of the factors that influence transition probability. Indeed, complex driving decisions are influenced by many factors including the attributes of the driver and characteristics of the context. More information about the biometrics of the driver and characteristics of the situation (vehicle kinematics and environmental variability) can paint a more complete picture of driving events. Having more variables can help with more in-depth analysis of driving behavior and hence a deeper understanding of driving decisions. Similar analysis with richer datasets containing the layout, speed and distances of surrounding vehicles as well as information about the weather, lighting, surface condition would probably yield better and more reliable results. Also, note that the results are obtained by aggregating different driver behaviors and the comparisons of proportions at the trip level. However, the MDP method provides more value when investigating individual behavior separately. Therefore, as more data from the same person becomes available, the proposed method will enable deeper learning about a person's driving behavior.

8. Conclusions

This study uses high frequency and diverse driving data to learn driving decisions. Specifically, reward and states are defined theoretically, and real-life driving decisions are analyzed in order to explore correlations with contextual factors i.e., proximity to surrounding objects. MDP is applied to learn personally revealed choices of drivers in terms of acceleration, deceleration and maintaining a constant speed. Eight states are defined based on the number of objects surrounding the host vehicle and distance to the front object. Individual driver reward functions are estimated using multinomial logit model. The results show that with increasing objects around a host vehicle, drivers would rather accelerate to avoid crowdedness around them.

Segmenting trips based on the level of crowdedness indicated that with increased level of crowdedness, fewer drivers choose acceleration as their PRCs because they are constrained to either keep constant speed or decelerate due to traffic condition. In addition, the level of following PRCs show that the obtained PRCs match drivers' real behavior 36% of the time.

One potential application of this study is to generate short-term predictive information about driver decisions, which can be used to warn the driver when they deviate substantially from their own PRCs based on their own historical driving performance. This can be implemented as a more intelligent ADAS, which can determine if the speed-up behind a slow-moving vehicle is due to driver impairment, or with a specific plan of overtaking it. Moreover, this determination need not be one-sided – disseminating a driver's preferred actions to surrounding vehicles may enable their drivers to foresee the states and actions of other drivers. An instrumented vehicle equipped with such model parameters can determine and anticipate the behavior of surrounding legacy vehicles and incorporate that information when selecting the most appropriate response and optimizing its

future trajectory. In order to develop applications, a considerable amount of trip data will be needed for each driver under different traffic conditions, which should not be a problem as instrumented vehicle big data are rapidly emerging.

Acknowledgement

This paper is based upon work supported by the U.S. National Science Foundation under Grant No. 1538139. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are very grateful to Mr. Zachary Jerome for editing help.

References

Arvin, R., Kamrani, M. & Khattak, A. J. 2019a. Examining the Role of Speed and Driving Stability on Crash Severity Using SHRP2 Naturalistic Driving Study Data. *Transportation Research Board 98th Annual Meeting.* Washington D.C.

Arvin, R., Kamrani, M. & Khattak, A. J. 2019b. The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data. *Accident Analysis & Prevention,* 132**,** 105226.

Bellman, R. 1954. The theory of dynamic programming. DTIC Document.

Brackstone, M. & Mcdonald, M. 1999. Car-following: a historical review. *Transportation Research Part F: Traffic Psychology and Behaviour,* 2**,** 181-196.

Burnham, G., Seo, J. & Bekey, G. 1974. Identification of human driver models in car following. *IEEE transactions on Automatic Control,* 19**,** 911-915.

Choudhury, C., Ramanujam, V. & Ben-Akiva, M. 2009. Modeling acceleration decisions for freeway merges. *Transportation Research Record: Journal of the Transportation Research Board***,** 45-57.

Cooper, R. A. 1991. System identification of human performance models. *IEEE transactions on systems, man, and cybernetics,* 21**,** 244-252.

Farah, H. & Koutsopoulos, H. N. 2014. Do cooperative systems make drivers' car-following behavior safer? *Transportation research part C: emerging technologies,* 41**,** 61-72.

Gipps, P. G. 1981. A behavioural car-following model for computer simulation. *Transportation Research Part B: Methodological,* 15**,** 105-111.

Gipps, P. G. 1986. A model for the structure of lane-changing decisions. *Transportation Research Part B: Methodological,* 20**,** 403-414.

Guzzella, L. & Sciarretta, A. 2007. *Vehicle propulsion systems*, Springer.

Hayeri, Y. M., Kim, K.-E. & Lee, D. 2016. An Inverse Reinforcement Learning Approach to Car Following Behaviors.

Henclewood, D. 2014. Safety Pilot Model Deployment – One Day Sample Data Environment Data Handbook. . *Research and Technology Innovation Administration, US Department of Transportation: McLean, VA*.

Hoogendoorn, R., Van Arem, B. & Hoogendoorn, S. Incorporating driver distraction in car-following models: Applying the TCI to the IDM. Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on, 2013. IEEE, 2274-2279.

Hoogendoorn, S., Hoogendoorn, R. G. & Daamen, W. 2011. Wiedemann revisited: new trajectory filtering technique and its implications for car-following modeling. *Transportation Research Record,* 2260**,** 152-162.

Kamrani, M. 2018. *INTEGRATING AND ANALYZING DRIVER, VEHICLE AND ROAD INFRASTRUCTURE VOLATILITIES USING CONNECTED AND INSTRUMENTED VEHICLES TECHNOLOGY.* Dissertation, The University of Tennessee, Knoxille.

Kamrani, M., Arvin, R. & Khattak, A. J. 2018a. Analyzing Highly Volatile Driving Trips Taken by Alternative Fuel Vehicles. *Transporation Research Board 97th Annual Meeting.* Washington DC.

Kamrani, M., Arvin, R. & Khattak, A. J. 2019. The Role of Aggressive Driving and Speeding in Road Safety: Insights from SHRP2 Naturalistic Driving Study Data. *Transportation Research Board 98th Annual Meeting.* Washington D.C.

Kamrani, M., Khattak, A. J. & Li, T. 2018b. A Framework to Process and Analyze Driver, Vehicle and Road infrastructure Volatilities in Real-time. *Transportation Research Board 97th Annual Meeting.* Washington DC.

Karan, F. S. N. & Chakraborty, S. 2016. Dynamics of a repulsive voter model. *IEEE Transactions on Computational Social Systems,* 3**,** 13-22.

Kiencke, U., Majjad, R. & Kramer, S. 1999. Modeling and performance analysis of a hybrid driver model. *Control Engineering Practice,* 7**,** 985-991.

Kiencke, U. & Nielsen, L. 2005. Vehicle modelling. *Automotive Control Systems: For Engine, Driveline, and Vehicle***,** 301-349.

Koppelman, F. S. & Bhat, C. 2006. A self instructing course in mode choice modeling: multinomial and nested logit models.

Koutsopoulos, H. N. & Farah, H. 2012. Latent class model for car following behavior. *Transportation research part B: methodological,* 46**,** 563-578.

Kuderer, M., Gulati, S. & Burgard, W. Learning driving styles for autonomous vehicles from demonstration. Robotics and Automation (ICRA), 2015 IEEE International Conference on, 2015. IEEE, 2641-2646.

Li, L., Chen, X. M. & Zhang, L. 2016. A global optimization algorithm for trajectory data based car-following model calibration. *Transportation Research Part C: Emerging Technologies,* 68**,** 311-332.

Liu, A. & Pentland, A. Towards real-time recognition of driver intentions. Intelligent Transportation System, 1997. ITSC'97., IEEE Conference on, 1997. IEEE, 236-241.

Macadam, C. C. 2003. Understanding and modeling the human driver. *Vehicle System Dynamics,* 40**,** 101-134.

Mohammadi, S., Kamrani, M., Khattak, A. J. & Chakraborty, S. 2019. Social Influence on Driver Decisions Using Modeling and Gossip Algorithms. *Transportation Research Board 98th Annual Meeting.* Washington D.C.

Ossen, S. & Hoogendoorn, S. P. 2005. Car-following behavior analysis from microscopic trajectory data. *Transportation Research Record,* 1934**,** 13-21.

Ossen, S. & Hoogendoorn, S. P. 2011. Heterogeneity in car-following behavior: Theory and empirics. *Transportation research part C: emerging technologies,* 19**,** 182-195.

Ossen, S., Hoogendoorn, S. P. & Gorte, B. G. 2006. Interdriver Differences in Car-Following: A Vehicle Trajectory–Based Study. *Transportation Research Record,* 1965**,** 121-129.

Papathanasopoulou, V. & Antoniou, C. 2015. Towards data-driven car-following models. *Transportation Research Part C: Emerging Technologies,* 55**,** 496-509.

Prokop, G. 2001. Modeling human vehicle driving by model predictive online optimization. *Vehicle System Dynamics,* 35**,** 19-53.

Ramming, M. S. 2001. Network knowledge and route choice. *Unpublished Ph. D. Thesis, Massachusetts Institute of Technology*.

Rothery, R. W. 1992. Car following models. *Trac Flow Theory*.

Salvucci, D., Boer, E. & Liu, A. 2001. Toward an integrated model of driver behavior in cognitive architecture. *Transportation Research Record: Journal of the Transportation Research Board***,** 9-16.

Sezer, V. 2018. Intelligent decision making for overtaking maneuver using mixed observable markov decision process. *Journal of Intelligent Transportation Systems,* 22**,** 201-217.

Shimosaka, M., Nishi, K., Sato, J. & Kataoka, H. Predicting driving behavior using inverse reinforcement learning with multiple reward functions towards environmental diversity. Intelligent Vehicles Symposium (IV), 2015 IEEE, 2015. IEEE, 567-572.

Sutton, R. S. & Barto, A. G. 1998. *Reinforcement learning: An introduction*, MIT press Cambridge.

Toledo, T., Koutsopoulos, H. N. & Ben-Akiva, M. 2007. Integrated driving behavior modeling.

Transportation Research Part C: Emerging Technologies, 15**,** 96-112.
Train, K. 1978. A validation test of a disaggregate mode choice model. *Transportation Research,* 12**,** 167-174.
Train, K. 1986. *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand*, MIT press.
Train, K. E. & Winston, C. 2007. Vehicle choice behavior and the declining market share of US automakers. *International economic review,* 48**,** 1469-1496.
Yang, Q. & Koutsopoulos, H. N. 1996. A microscopic traffic simulator for evaluation of dynamic traffic management systems. *Transportation Research Part C: Emerging Technologies,* 4**,** 113-129.
Ye, Y., Zhang, X. & Sun, J. 2019. Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment. *Transportation Research Part C: Emerging Technologies,* 107**,** 155-170.

List of Keywords

Driving Behavior
Markov Decision Processes
Basic Safety Messages
BSM
Multinomial Logit Regression
Connected Vehicle Data

List of Acronyms

BSM: Basic Safety Messages. Information exchanged between connected vehicles at approximately 10x per second that contain data elements such as position, speed, acceleration, heading and etc.

CAV: Connected and Automated Vehicle. A vehicle that is capable of sensing its environment and moving safely with little or no human input as well as sending and receiving information from other vehicles, infrastructure, bikes, pedestrian etc.

DAS: Data Acquisition Systems. A collection of software and hardware that allows one to measure or control physical characteristics of something in the real world

MDP: Markov Decision Process. A mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker.

MNL: Multinomial Logit Regression Model. The regression analysis to conduct when the dependent variable is nominal with more than two levels.

PRC: Personally Revealed Choices. Equivalent to Optimal Policy in Markov Decision Process framework and that is the solution for an MDP which describes the best action for each state in the MDP.

RL: Reinforcement Learning. An area of machine learning concerned with how an agent ought to take actions in an environment in order to maximize some notion of cumulative reward.